Structuring and Fusing Text

Dr. David F. Noble
Evidence Based Research, Inc.
1595 Spring Hill Road, Suite 250
Vienna, VA 22182
USA

noble@ebrinc.com

ABSTRACT

Much information important for battlefield assessments is transmitted as unstructured text, including both unclassified documents available to the general public as well as highly classified reports and messages. Text can convey such highly valued information as adversary plans and goals. Unfortunately, valuable text nuggets may be buried in massive amounts of less important information, and may be difficult to find. Once found, different aspects of an entity or activity may be scattered among different text sources, making it difficult to assemble into a coherent picture able to convey context, relationships, and trends. This paper describes a methodology for structuring and fusing open source information so that it may be presented on maps and diagrams. This process employs formal ontologies for the fusion domain and for evidential reasoning, a commercial tool for text extraction and structuring, and tools to help operators review, edit, and augment the fusion products. It features a fusion pedigree to document the audit trail of sources and processes contributing to a fusion product. The fusion steps are: 1) collect structured and unstructured information related to the entity or event of interest; 2) extract and structure free text; 3) create "event reports" from each structured record, whether derived from free text or previously structured sources; 4) create ontology-based communication reports from these records, 5) associate these reports, 6) cue manual search for additional information, and 7) fuse the information.

1.0 ADVANTAGES FROM STRUCTURING AND FUSING TEXT

Unstructured text is a pervasive medium for communicating information. Newspapers, periodicals, messages, internet pages, and reports are all primarily text. The text may be open source, available to the general public; or may be highly classified messages and reports. In addition to its wide availability, text is a very flexible representation medium, able to convey anything that people are capable of discussing, from simple descriptions of objects and activities to abstract scientific theories. Text can communicate all types of fusion products from level 1 descriptions of entities and activities to level 2 descriptions of organizations and relationships and level 3 descriptions of goals and intent. Text information can also very timely. The huge proliferation of open source information on the Internet, including news sites, discussion boards, and chat rooms, often provides the initial reporting and early indicators of important events and activities. [1]

Unfortunately, understanding information conveyed as text can require considerable work. It's hard to find key information if doing so requires reviewing thousands of documents. It can be hard to find significant trends and relationships in text when the elements of these relationships are buried in widely scattered places [2]. Often, graphical depictions are a much more efficient way to convey information than text is. People

Noble, D.F. (2006) Structuring and Fusing Text. In *Information Fusion for Command Support* (pp. 10-1 – 10-14). Meeting Proceedings RTO-MP-IST-055, Paper 10. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.asp.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 DEC 2006	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Structuring and Fusing Text	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Evidence Based Research, Inc. 1595 Spring Hill Road, Suite 250 Vienna, VA 22182 USA**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002031., The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**32**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

often can understand information much more readily when that information is presented on maps, charts, and graphs than when it's presented as text.

This paper describes how to process information in text so that it can be consolidated, converted into well-defined structured records, and presented graphically on maps, charts, and diagrams. Such graphical presentations help people focus on the most important information in the text, and help them understand relationships and trends. Because fusion manages uncertainty, text fusion can process the uncertain information found in various sources creating consolidated information records that are more complete, accurate and precise than the information in any single source. This representation of uncertainty helps people understand both what is not known as well as what is, and help them to hedge for these uncertainties.

This paper addresses fusion in two ways. First, it describes how to fuse text. It describes a method for finding partial descriptions of an activity or entity in multiple sources, and then describes how to put these descriptions together into a more complete, precise, and accurate single description. Though the products of text fusion can be anything, including such Level II and III products as adversary plans and goals, the text fusion processes and models themselves are those familiar in Level 1 fusion. Second, this paper is also about an alternative method of obtaining level II and III fusion products: by combining partial level II and III descriptions in text and processing the text to create a more complete and accurate Level II or III fusion product.

2.0 EXAMPLE OF OPEN SOURCE FUSION

The following example illustrates the conversion of the information in multiple unstructured sources into a single fused structured record. The example is based on the work EBR is now performing for a client. We have modified the material somewhat, both to simplify the example and to make it more instructive.

In this case, the client is interested in understanding commercial relationships among companies within the telecommunications industry. Among their interests are various technical and marketing associations. In this example, there are two sources, PWID 1 (Published Works #1) and PWID 2. Each source contains multiple references to the nature of the association, with each reference contributing some information.

PWID 1

Indonesia's Telkom awards Ericsson with broadband contract

Xinhua, 11/11/2004 15:01

Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services. Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.

"There is already a very small number of lines—lower than 1, 000—available in Surabaya," PT Ericsson Indonesia President Mitch Lewis was quoted as saying.

He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April.

Telkom launched its broadband service, called Speedy, in July and is aiming to provide 40,000 connections in Jakarta, and 10,000 in Surabaya in the first phase of the project.

It has allocated US$15 million for the service's infrastructure, or between US$290 and US$300 per line.

Figure 1a: Text Input to Fusion Process (PWID 1)

PWID 2

Ericsson wins deal in Indonesia

11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers.

The broadband service will include DSL access that will deliver data, voice, and video simultaneously in East Java. Ericsson's work for this contract is slated to begin in April of 2005.

PT Telkom is Indonesia's largest telecommunications company and currently offers a broadband service, called Speedy, which it launched in July of 2004. State-run Telkom plans to convert its network's infrastructure broadband capacity from fewer than 1,000 lines to 2 million lines by 2008.

Figure 1b: Text Input to Fusion Process (PWID 2)

The fusion process extracts and combines the information in these two sources to create the following single record (Figure 2):

Name	
OrgAssocType	Technology Partnership
OrganizationsAndRoles (1)	
Organization	Ericsson
Role:	Vendor
OrganizationsAndRoles (2)	
Organization	PT Telkom
Role:	Buyer
AssocStartDate	
Min estimate	April 1, 2004
Max estimate	April 15, 2004
Confidence	High
LocationOfWork	
City	Surabaya
Region	East Java
Nation	Indonesia
Technology	Broadband
Contract Amount	
Estimate	$7.5 million
Lower bound	$7.5 million
Upper bound	$7.5 million
Confidence	High
Communication Reports	Report 1; Report 2; Report 3, Report 4

Figure. 2: Product of Open Source Fusion

All of the information in the structured record can be found dispersed within the two sources depicted in Figure 1. However, unlike the text representation of the information, the representation as a structured record may be readily combined with additional sources of information, both structured and unstructured, and may be depicted graphically with other information to show trends and relationships.

Interestingly, based on the types of models needed, the fusion of open source information seems to be Level 1 fusion as defined by the Joint Directors of Laboratories Fusion Panel. As will be described, this fusion process entails the usual steps of level 1 fusion: obtain sensor reports, align report data, condition sensor reports using sensor models, estimate association probabilities among reports, associate applying strategy for managing uncertain associations, and fuse and update state [3], [4], [5].

In addition, open source fusion requires a "collection management" process in which more coarse-grained broad coverage sensors cue more fine-grained narrow coverage ones. In our case, the course grained sensor is Aerotext™, a state of the art text extractor from Lockheed Martin. Aerotext™ can search vast amounts of information quickly, extracting and structuring with a precision that greatly facilitates the information fusion process. The fine-grained sensors are people who cannot search quickly but can extract many more details from the text than the automated text extractor can. Accordingly, our system helps operators to review and edit the fusion products generated at each of the processing steps. This not only improves the quality of the product, but the process's transparency and user control help build user confidence in the fusion products [6].

3.0 FUSION ENVIRONMENT AND TOOLS

The fusion process to be described requires an information processing infrastructure for collecting, structuring, and managing information. Figure 3 depicts the major components of the system that EBR uses to collect and structure open source information [7]. At EBR we use primarily Kapow™ to collect open source information, converting Web pages into plain text. We then use Lockheed Martin's AeroText™ to structure the information in the text, creating structured records with all fields able to be mapped to terms of a formal ontology.

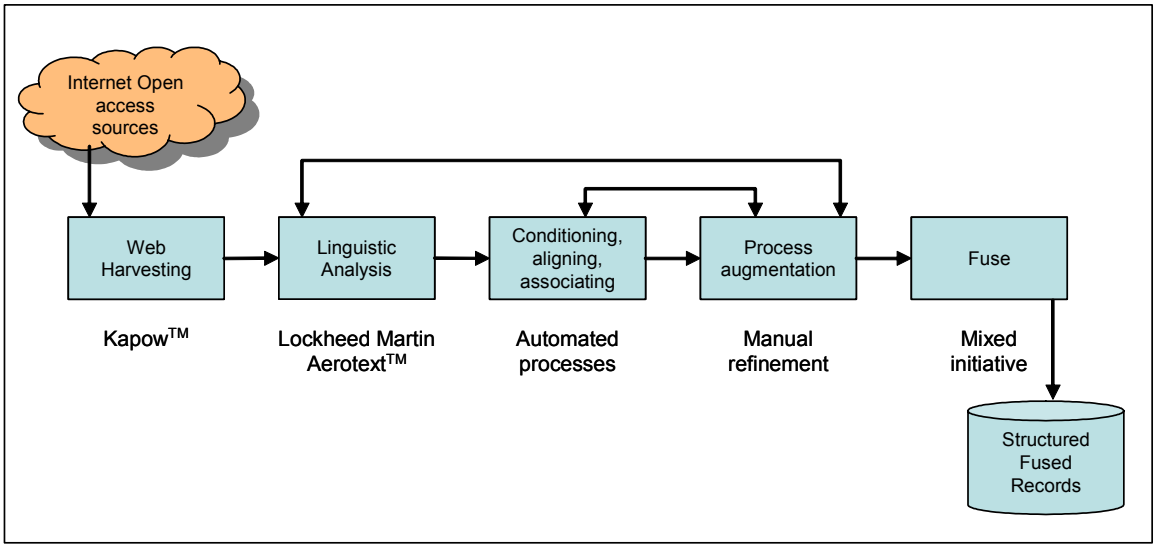


Figure 3: Open source collection, structuring, and fusion tools

The immediate product of the text extraction and structuring are Aerotext™ "events." These events are analogous to raw sensor reports. These raw reports need to be converted into more robust ontology-defined reports, more suitable for fusion. These reports are analogous to the track reports in tracking. To create these more robust reports, the Aerotext™ output needs to be expanded, aligned and conditioned. Here, the

expansion is incorporation of additional information available in the text but not extracted by Aerotext™. The alignment is enforcement of a common vocabulary defined in a formal ontology terminology list. The conditioning is association of uncertainties with some of the data fields. These uncertainties then assist in association decisions.

While the goal of the fusion methodology is full automation, at the present time the process requires considerable manual intervention. All of the expansion is currently manual. Manual refinement improves conditioning, alignment, combination, and state estimation.

The remainder of this paper will illustrate each of these steps for the problem presented in section 2.

4.0 OPEN SOURCE FUSION STEPS

This paper describes seven steps in open source fusion: 1) process set up, 2) information collection and extraction; 3) conversion of Aerotext™ events into ontology-defined aligned and conditioned "communication" records, 4) manual refinement of the communication records, 5) association of these records; 6) cued manual search for further refining information in the source text, and 7) fusion and state update.

4.1 Process set-up

Process set-up is the work required to prepare the system for fusion. It includes defining the ontology for the domain being examined, setting up Aerotext™ to extract the information defined in the ontology, and generating alignment rules.

In its open source collection and structuring work, EBR employs an ontology [8] based on the Suggested Upper Merged Ontology (SUMO), an effort within the IEEE SUO working group to create a high level ontology for use by expert systems within a variety of domains [9]. We use the OWL file structure and the Protégé application to create and edit the ontology.

Our adaptation of SUMO has four principal components: people and organizations, competitive intelligence, telecommunications, and evidential reasoning. The record shown in Figure 2 is a reflection of the "Organization Association" class in the ontology. The ontology-defined analogues of track reports are the "Communication" classes. The subclass employed in this article is called "Organization Association-Communication." These communication classes are the buffer between the immediate Aerotext™ output (which are not constrained by the ontology) and the final fusion products (like the "OrganizationAssociation" records) which analysts examine to understand the collected information.

The system's ontology serves defines the schema for the relational database that stores fusion products. Specialized software generates the database schema from the ontology. This specialized software can extend the database schema to accommodate extensions to the ontology within a few minutes. It can also generate new user database review and input forms within a few minutes. This design provides the foundation for system agility—its ability to quickly address new domains.

The second step in the process set-up is creating the Aerotext™ extraction rules. These rules describe for Aerotext™ how to identify and structure the information to be extracted. In effect, they create fairly abstract templates that describe all the different ways that a concept can be expressed in the target language. In our example, these are the different ways that one can express in English that two organizations are entering into

an association. Adapting rules to accommodate new issues in a previously examined domain can often be accomplished quickly. Developing rules for a new domain can be labor intensive, sometimes requiring more than a month of effort from experienced Aerotext™ users.

The Aerotext™ rules that we use specify what to look for within an English sentence and how to represent the information found as structured event records. That is, the rules tell Aerotext™ how to recognize the issues of interest to be found in free text, and how to deposit the information contained in a sentence into a structured record.

Unfortunately, Aerotext™ currently has significant limitations that prevent its use as a commercial off the shelf fusion engine. First, though it can unify some information across sentences, it cannot unify the information needed for comprehensive battlefield assessments. Second, Aerotext™ does not understand much about grammar. It cannot easily differentiate between the subject and objects of sentences, and thus has difficult in filling such data fields as the company "roles" in Figure 2 because it cannot distinguish the "buyer" from the seller in the sentences shown in Figure 1.

The third issue in set-up is describing to the system how it is to translate the terms in the Aerotext™ output records into the ontology-defined terms in the fusion database. This is required because the vocabulary demands of Aerotext™ for text extraction and of fusion process for well-defined concepts conflict. Aertotext™ wants to know all the different words that can be used to express an idea because that helps it detect and structure these ideas. For example, it needs to know the many different ways that text may refer to an organization. The fusion process, on the other hand, wants to use only a single term to describe a concept, because it wants to ensure that, to the extent possible, when words in the database are different, they define different entities or events, and when they are they same, they define the same entities and events. Thus, the fusion process needs to use just one way to refer to each organization. This vocabulary alignment is very helpful in the association step of fusion for it helps the system judge whether or not two extracted text records are referring to the same or different things.

4.2 Information collection and extraction

Once the set-up is completed, the system is ready to begin its fusion of open source information. The first step is to collect material that may contain useful information, and then to identify, extract, and structure the information. As indicated in Figure 3, EBR's principal tool for automated collection is Kapow™. Kapow uses intelligent spider technologies to auto-harvest / extract information of interest off the web and then saves this information in a database. The next step is for Aerotext™ to find and extract the relevant information. In this case, there are five sentences that contain the information that contribute to the final fused product shown in Figure 2. As noted in Figure 4, Aerotext™ finds four of them, and converts these four to the structured Aerotext event records.

Aerotext™ labels each of the extracted sentences with an identifier, consisting of the published works ID and the region within the published work. The region is a binary (a,b) where "a" is the number of "tokens." (basically, words) of the first word from the beginning of the source and the "b" is the number for the last word in the extraction. As will be shown, this location labeling is very important for associating events with each other.

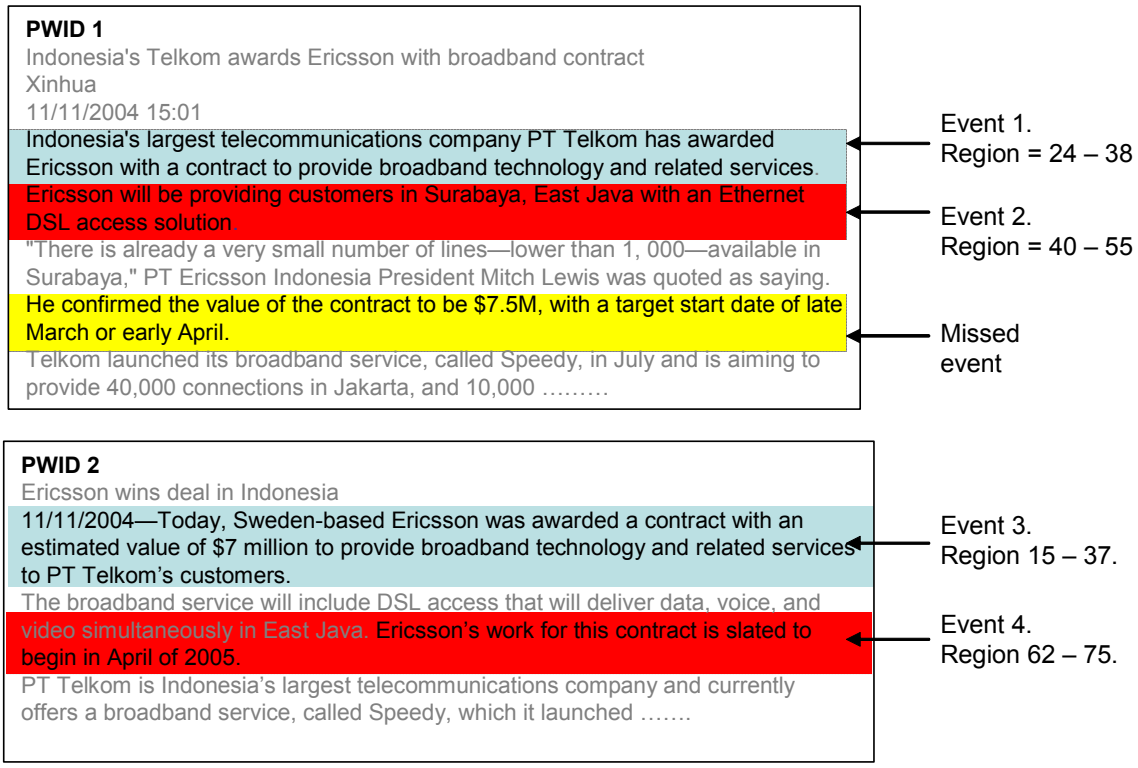


Figure 4: Information Extracts

Aerotext converts each of these extractions into a structured event record, as shown in Figure 5.

	Event 1	Event 2	Event 3	Event 4
Source	PWID 1	PWID 1	PWID 2	PWID 2
Region	24, 38	40, 55	15, 37	62, 75
Event type	Contract Event	Technology Event	Contract Event	Contract Event
Text	Telkom has awarded Ericsson with a contract to.....	Ericsson will be providing customers in Surabaya, East Java...	Ericsson was awarded a contract with an estimated value of $7 million to ...	Ericsson's work for this contract is slated to begin in April of 2005
Subtype	Technology		Technology	General
Organization 1	Telkom	Ericsson	Ericsson	Ericsson
Organization 2	Ericsson		PT Telkom	
Contract_Place		Surabaya, East Java		
Contract_Amount			$7 million	
Contract_Date				in April of 2005
Technology	broadband	Ethernet, DSL	broadband	

Figure 5: Structured Aerotext™ Records

Note that in our use of Aerotext™, the unit of extraction is one sentence. Therefore, information that spans multiple sentences will be represented in multiple structured events.

In this example, there are six desired pieces of information about the association: 1) the organizations involved; 2) the roles of each of the organizations; 3) the date the association begins; 4) the location of the work; 5) the size of the award; and 6) the technology involved (broadband). Each of the extracted sentences has some of the needed information. The first and third records state both of the organizations and their roles, the association date is in event 4 (and the missed event), the location is in event 2, the size of the award is in event 3 (an estimate) and in the missing event (confirmed amount), and the technology is in events 1, 2, and 3. In addition, the missing event, after it is found, can refine the estimates of contract size in event 3 and start date in event 4. The purpose of the fusion is to create a single record (Figure 2) that combines all of these data and also refines the award amount and start date.

Aerotext™ in finding and structuring these sentences provides an enormously important function, because it can quickly process very large amounts of material that would be uneconomical for people to do. However, comparing the initial material with the text shows some of the limitations of our extraction (better rules would extract better, but they take time to create). First, Aerotext™ missed one of the key sentences. It lacked the specific information that our rules required in order to find an event. Second, it was unable to distinguish between the organization roles of PT Telkcom and Ericcson, since the rules didn't specify how to distinguish between vendor and buyer. Third, it used two different labels for PT Telkom: "PT Telkom" and "Telkom." Fourth, the third event did not note that the value of the contract was "estimated." Fifth, the contract start date in event 4, "in April of 2005" remained a string rather than being interpreted as a date. The next two steps, which convert the Aerotext™ events into ontology-defined "communications" records, address all except the first of these issues.

4.3 Conversion to "communication" records

This step starts the conditioning and alignment of the Aerotext™ events. It is the step that sets the stage for associating the five records (after the fifth is found) and then creating a single fused record that contains all of the data. Also, as noted previously, this step is analogous to creating track reports from sensor data. Figure 6 shows the "communication" records created from the first two event records of Figure 5. The communication record is defined by the ontology (unlike the Aerotext™ events), and closely resembles the fusion product "organization-association" record of Figure 2. The system automatically created these two records. Other than reformatting the data in a structure more suitable for fusion, the conversion to "communication" records did not entail much processing. The system did some alignment, changing "Telkom" in the Aerotext™ event to "PT Telkom," and parsing the location of the work. Both of these alignments will help with the association. The system was not, however, able to do automated conditioning. It did not, for example, attempt to define the level of uncertainty in the date estimate (in report 4) or size of the contract. Moreover, this automatic processing did not address most of the problems noted in Section 4.2 This will require operator intervention.

Name	Report 1	Report 2	Report 3	Report 4
OrgAssocType				
Organizations(1)				
Organization	PT Telkom	Ericsson	Ericsson	Ericsson
Role:				
Organizations (2)				
Organization	Ericsson		PT Telkom	
role				
AssocStartDate				
Lower est.				04/01/2005
Upper est.				
LocationOfWork: City Providence Country		 Surabaya East Java Indonesia		
Technology	Broadband	Ethernet, DSL	broadband	
Contract Amount			$7 million	
Estimate				
Lower bound				
Upper bound				
Confidence				
Pedigree				
Text	Telkom has ...	Ericsson will be	Ericsson was	Ericsson's
Processor	Aerotext	Aerotext	Aerotext	Aerotext
Source ID	PWID 1	PWID 1	PWID 2	PWID 2
Region	24, 38	40, 55	15, 37	62, 75

Figure 6: Ontology-based Communication Records

4.4 Manual refinement of the communication records

The Aerotext™ rules that EBR developed for organization-association events are not detailed enough to extract the organization role. Given the current state of the art in automated text extraction, the EBR Aerotext™ team did not feel that it was economical to generate the more complicated rules that might be able to extract this information. Accordingly, EBR instead is creating an environment to facilitate manual entry of the additional data. This manual entry not only improves the quality of the fusion products, but also helps the user understand and control the fusion process, an important contributor to sustaining user confidence. EBR is also considering supplementing Aerotext™ with a parser to provide this information.

Figure 7 summarizes the communication records after operator editing. Operator changes are in italics. Note that the operators inserted the organization roles, easily found in the first and third events. Operators also represented the two uncertainties. In order to maintain an audit trail describing the evolution of the data, the operators who updated the records inserted their initials into the pedigree information.

Name	Report 1	Report 2	Report 3	Report 4
OrgAssocType				
Organizations(1)				
Organization	Ericsson	Ericsson	Ericsson	Ericsson
Role:	*Vender*		*Vender*	
Organizations (2)				
Organization	PT Telkom		PT Telkom	
Role	*Buyer*		*Buyer*	
AssocStartDate				
Lower est.				*04/01/2005*
Upper est.				*04/30/2005*
LocationOfWork: City Providence Country		 Surabaya East Java Indonesia		
Technology	Broadband	Ethernet, DSL	broadband	
Contract Amount				
Estimate			*$7 million*	
Lower bound			*$5 million*	
Upper bound			*$9 million*	
Confidence			*moderate*	
Pedigree				
Text	Telkom has ...	Ericsson will be	Ericsson was	Ericsson's work
Processor	Aerotext, AJJ	Aerotext AJJ	Aerotext, AJJ	Aerotext, AJJ
Source ID	PWID 1	PWID 1	PWID 2	PWID 2
Region	24, 38	40, 55	15, 37	62, 75

Figure 7: Ontology-based Communication Records

4.5 Associate reports

Deciding whether reports are actually referencing the same incident is the report association problem, which for many fusion problems is the principal challenge. Reports that refer to the same entity or activity can be combined to estimate more accurately the characteristics of the entity or activity. Reports that reference separate activities or entities cannot be directly fused. When two reports may or may not be about the same activity or entity, then the fusion logic needs to decide what to do. It can tentatively combine them, with the provision that the reports can be disassociated if necessary, can set them aside with the provision that they can be combined later, or can generate "multiple hypotheses" in which the reports are combined in one hypothesis and not the other. The latter can lead to very complicated hypothesis management logic.

In fusion of open source information, association decisions are based on two criteria: 1) the compatibility and similarity of the content of the data in the reports and 2) the proximity of the text "regions" of the source material.

As summarized in Figure 8, in this example, Reports 1 and 3 are associated based on content similarity and compatibility. They are both about "broadband" association with the same organizations in the same role and appeared in news reports on the same day. This amount of match is sufficient to associate the reports, even though they are drawn from different sources.

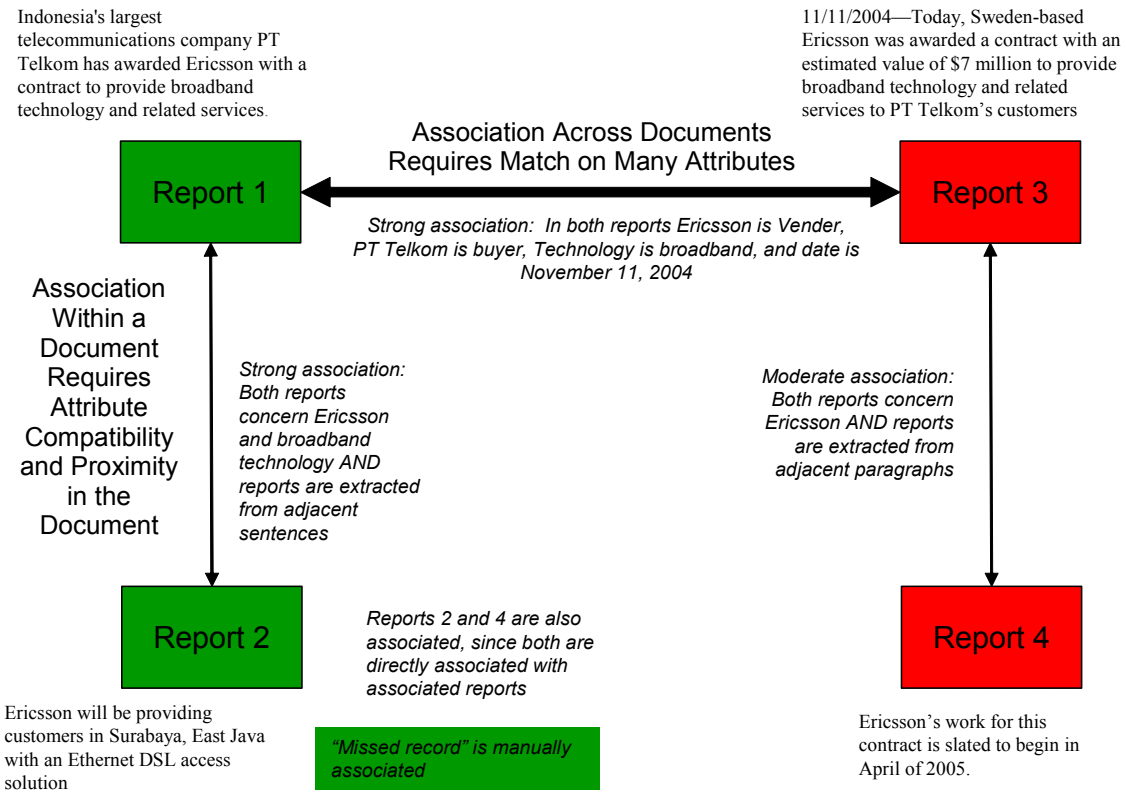


Figure 8: Report Associations

In contrast, reports 1 and 2 cannot be matched based on the similarity of content. They are compatible, since both concern Ericsson and broadband technology ("broadband" and "Ethernet DSL" do not conflict because our ontology knows that Ethernet DSL is a kind of broadband, but an operator can also make this judgment). This degree of match by itself would not be sufficient to associate the reports. However, these two reports are adjacent sentences in the original text, as is determined from the source ID and region fields in the pedigree part of the reports. Because adjacent sentences frequently discuss the same subject, the fact that report 2 directly followed report 1 in the text and the compatibility of the content is sufficient to make the match. Similarly, reports 3 and 4 are matched.

EBR has not yet formalized general rules for report association, and will be investigating proximity measures. This proximity would need to reflect the probability that two reports reference the same events and entities given the degree of similarity in their fields. Establishing general criteria for report association, however, a very significant part of developing this fusion methodology, for incorrect associations can completely invalidate conclusions based on the association. Such events occasionally get into the news, as was the case when an infant with the same name as someone on a terrorist watch list was not permitted to board a plane.

In the meantime, we plan to defer associating reports that could plausibly be about separate events or entities. This prevents the kind of embarrassing mistake cited above.

4.6 Cue manual search

In examining the reports on the Ericsson–PT Telkom association, an operator might note that the size of the contract is not confirmed, but is only an estimate. Knowing that the automatic text extraction can miss important material, an operator might decide to manually inspect the source material that the other reports are drawn from. As noted previously, this is analogous to a broad view sensor cuing a fine-grained one in surveillance, with in this case Aerotext™ serving as the broad view sensor and people serving as the fine grained one. Once a person reads the source material in "published works 1," he or she will note the missed sentence:

"He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April."

The operator can incorporate this material into the process by manually creating an association communication record. He can specify that the contract amount is $7.5M, and that this number has very high confidence. He can also note that the start date has an early estimate of March 15 and a late estimate of April 15.

4.7Fusion and state update

The final step in this example is to combine the information in the five records (including the manually produced one) into a single record. In this example, there are three combination rules: 1) if all reports agree on a value, then put that value in the fused product; 2) if some reports cite a value that others do not contradict, then insert those values; 3) if reports reference compatible values, insert the most specific one the ontology uses; 4) if reports cite different numerical values with uncertainty estimates, then update using established state estimation methods.

The result of applying these fusion rules is the fused product shown in Figure 2.

As part of its "evidential reasoning" ontology, EBR has included several different ways for representing uncertainty. One of the most useful for open source fusion is a probability list of nominal data. Thus, for example, if two reports give different estimates of an entity identity, the fused product is just a list of these two estimates, each associated with a probability.

The system also permits analysts to comment on the trustworthiness and credibility of any data field.

5.0 CONCLUSION

Because of the prevalence and information value of text, the payoff from integrating text-derived data into graphical situation pictures is very high. In theory, it should be possible to use information available from text to enrich a situation display with critical information about adversary capabilities, plans, and intent. In practice, this is hard to do because of the difficulty of finding key information, extracting and structuring free text, and combining non-numeric information. Today, new tools make it feasible to efficiently collect and structure the needed information. As described in this paper, these same tools now also make it possible to combine information from multiple sources to create a much more complete and precise information product.

6.0 REFERENCES

[1] David Noble "Assessing the Reliability of Open Source Information" Proceedings of the 7th International Conference on Information Fusion. Stockholm, Sweden, 2004.

[2] Sarah Taylor, "Improving Analysis with Information Extraction Technology." Proceedings of the 8th International Command and Control Research Symposium. National Defense University, Washington, D.C., 2003.

[3] Edward Waltz and James Llinas. Multisensor Data Fusion. Artech House Publishers, 1990.

[4] John Robusto James Llinas, & David Noble. "Joint Exploitation Module." In Proceedings of the 1994 Tri-Service Data Fusion Symposium. Applied Research Laboratory. Laurel, Maryland. 1994.

[5] Tom Cool & David Noble. "Intelligence and Object Data Base Generator Tools." In Proceedings of the 1994 Tri-Service Data Fusion Symposium. Applied Research Laboratory. Laurel, Maryland. 1994.

[6] Erik Blasch. "Situation, Impact, and User Refinement. In Signal Processing, Sensor Fusion, and Target Recognition XII. Edited by Kadar, Ivan. Proceedings of the SPIE, Volume 5096, pp. 463-472 (2003)

[7] Steve Shaker and Victor Richardson. "Putting The System Back Into Early Warning," Competitive Intelligence Magazine; May-June 2004; pp. 13-17, 2004.

[8] M. M. Kokar, C. J. Matheus, K. Baclawski, J. A. Letkowski, M. Hinman, and J. Salerno. "Use Cases for Ontologies in Information Fusion." In Proceedings of the Seventh International Conference on Information Fusion, pages 415–421, 2004.

[9] L. Niles, and A. Pease. "Toward a Standard Upper Ontology" in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems.

Structuring and Fusing Text

Presented by

David Noble

Evidence Based Research

Presented at

North Atlantic Treaty Organization
Research and Technology Organization Specialists Meeting
Information Fusion for Command Support
The Hague Netherlands

November 9, 2005

Agenda

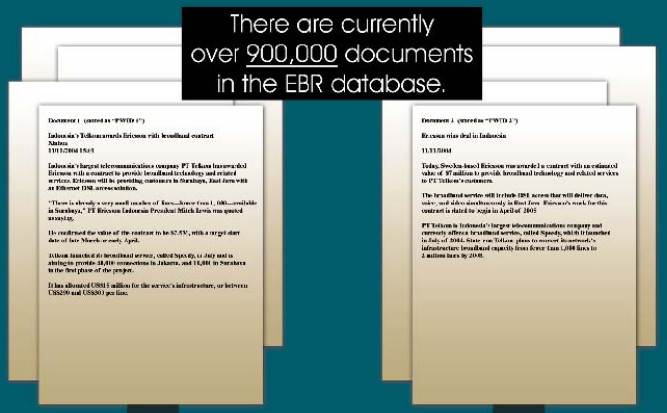
- Text Fusion
- Processing Steps
- Challenges

Text Fusion: Converting Unstructured Text to Graphical Data

FROM HUGE VOLUMES OF UNDIGESTED TEXT

Step 1: Collect Information

The system collects massive amounts of unstructured information from the Internet and other sources, leveraging Kapow™ automated collection resources.

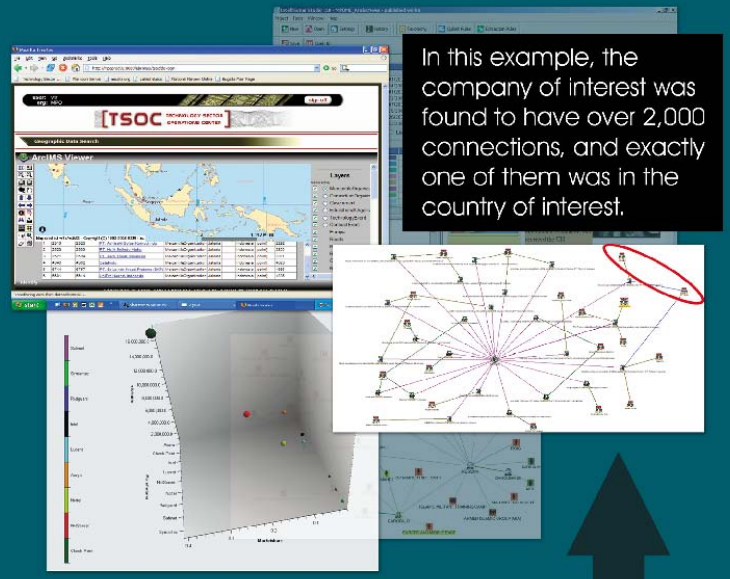


Step 2: Find Sentences

TO EASY-TO-REVIEW GRAPHS AND DIAGRAMS

Step 7: Visualize Structured Information

Once the information is structured into well-defined records, analysts use commercial visualization tools to analyze the information, find trends, compare viewpoints, document relationships, and discover the unexpected.



Find & Tag → Extract → Align → Condition → Fuse → Plot

Operational Benefits

- Reduce information overload
- Graphical representation of concepts
- Augmentation of C2 operational picture with level III products obtained from text
- Improve completeness and precision of operational pictures
- Improve shared understanding of operational situation

Example

Convert Text Documents such as These

PWID 1
Indonesia's Telkom awards Ericsson with broadband contract
Xinhua
11/11/2004 15:01
Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services. Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.
"There is already a very small number of lines—lower than 1,000—available in Surabaya," PT Ericsson Indonesia President Mitch Lewis was quoted as saying.
He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April.
Telkom launched its broadband service, called Speedy, in July and is aiming to provide 40,000 connections in Jakarta, and 10,000 in Surabaya in the first phase of the project.
It has allocated US$15 million for the service's infrastructure, or between US$290 and US$300 per line.

PWID 2
Ericsson wins deal in Indonesia
11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers.
The broadband service will include DSL access that will deliver data, voice, and video simultaneously in East Java. Ericsson's work for this contract is slated to begin in April of 2005.
PT Telkom is Indonesia's largest telecommunications company and currently offers a broadband service, called Speedy, which it launched in July of 2004. State-run Telkom plans to convert its network's infrastructure broadband capacity from fewer than 1,000 lines to 2 million lines by 2008.

To Plottable Data Records Such as This

OrgAssocType	Technology Partnership
OrganizationsAndRoles (1)	
Organization	Ericsson
Role:	Vendor
OrganizationsAndRoles (2)	
Organization	PT Telkom
Role:	Buyer
AssocStartDate	
Min estimate	April 1, 2004
Max estimate	April 15, 2004
Confidence	High
LocationOfWork	
City	Surabaya
Region	East Java
Nation	Indonesia
Technology	Broadband
Contract Amount	
Estimate	$7.5 million
Lower bound	$7.5 million
Upper bound	$7.5 million
Confidence	High
Communication Reports	Report 1; Report 2; Report 3, Report 4

Steps for Text Fusion

Fusion Step	As Applied in Text Fusion
Find and Tag	Find and tag sentences of interest in text documents
Extract	Convert text to structured records, using commercial extraction tools
Align	Convert terms to standard ontology-defined vocabulary and use ontology-defined records
Condition	Represent data precision as stated in text Estimate and represent data credibility
Associate	Decide which reports reference the same entity or activity
Fuse	Create / update consolidated records refining values in contributing structured records

Step 1: Find and Tag

PWID 1

Indonesia's Telkom awards Ericsson with broadband contract
Xinhua
11/11/2004 15:01

Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services. Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.

"There is already a very small number of lines—lower than 1, 000—available in Surabaya," PT Ericsson Indonesia President Mitch Lewis was quoted as saying.

He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April.

Telkom launched its broadband service, called Speedy, in July and is aiming to provide 40,000 connections in Jakarta, and 10,000 in Surabaya in the first phase of the project.
It has allocated US$15 million for the service's infrastructure, or between US$290 and US$300 per line.

PWID 2

Ericsson wins deal in Indonesia

11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers.

The broadband service will include DSL access that will deliver data, voice, and video simultaneously in East Java. Ericsson's work for this contract is slated to begin in April of 2005.

PT Telkom is Indonesia's largest telecommunications company and currently offers a broadband service, called Speedy, which it launched in July of 2004. State-run Telkom plans to convert its network's infrastructure broadband capacity from fewer than 1,000 lines to 2 million lines by 2008.

Document tags: telecom, broadband, Indonesia

Event 1. Region = 24 – 38

Event 2. Region = 40 – 55

Missed event

Event 3. Region 15 – 37.

Event 4. Region 62 – 75.

- Internet crawlers and semantic analyzers find and tag articles of interest
- Text extractors find specific sentences

Step 2: Extract

PWID 1

Indonesia's Telkom awards Ericsson with broadband contract

Xinhua

11/11/2004 15:01

Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services. Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.

"There is already a very small number of lines—lower than 1, 000—available in Surabaya," PT Ericsson Indonesia President Mitch Lewis was quoted as saying.

He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April.

Telkom launched its broadband service, called Speedy, in July and is aiming to provide 40,000 connections in Jakarta, and 10,000 in Surabaya in the first phase of the project.

It has allocated US$15 million for the service's infrastructure, or between US$290 and US$300 per line.

	Event 1
Source	PWID 1
Region	24, 38
Event type	Contract Event
Text	Telkom has awarded Ericsson with a contract to.....
Subtype	Technology
Organization 1	Telkom
Organization 2	Ericsson
Contract_Place	
Contract_Amount	
Contract_Date	
Technology	broadband

EBR uses Lockheed Martin's Aerotext for initial structuring. Some fields require additional manual editing.

Extraction State of Art

- Entity extraction finds entities (people, technologies, organizations, etc.) of interest. Not sufficient for text fusion
- Activity/ event extraction relates actions, actors, roles, time and qualifiers. Centers on verbs. Needed for text fusion
- State of the art for activity extractors is improving rapidly
- Correctly identifying event types and elements depends on ability to exploit grammar and semantics in natural language
- Extraction and structuring can require considerable analyst set up time. Approaches that exploit syntax and semantics in language require less time
- Human-level understanding requires extensive “common sense knowledge,” unlikely to be available in commercial tools soon

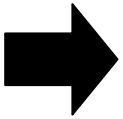
Hierarchy of Extraction Capabilities*

	Named Entity Recognition	Part of Speech Tagging	Sentence Structure Tagging	Multi-sentence Processing
Supplemental capabilities	Extract entities Extract entity patterns	Appositive handling Identify verb tense and voice	Syntactic role recognition Modifier phrase attachment	Sentence linking Pronoun resolution Tense matching
Finding sentences of interest	Worst →		→ Better	Much better
Multi-sentence processing	←	Some entity resolution	→	Multiple sentences but only within same documents
Analyst set up time	Highest →			→ Lowest
Role assignment	Difficult	Difficult	Easy	Easiest
Qualifier phrase linking	Very limited	Limited	Full	Full
Formally interpret uncertainty	No	No	No	No

* Adapted from “Natural Language Processing and Text Extraction,” Attensity

Step 3: Align

	Event 1
Source	PWID 1
Region	24, 38
Event type	Contract Event
Text	Telkom has awarded Ericsson with a contract to.....
Subtype	Technology
Organization 1	Telkom
Organization 2	Ericsson
Contract_Place	
Contract_Amount	
Contract_Date	
Technology	broadband



Name	Report 1
OrgAssocType	
Organizations(1) Organization Role:	PT Telkom
Organizations (2) Organization Role	Ericsson
AssocStartDate Lower est. Upper est.	
LocationOfWork: City Providence Country	
Technology	Broadband
Contract Amount Estimate Lower bound Upper bound Confidence	
Pedigree Text Processor Source ID Region	Telkom has ... Aerotext PWID 1 24, 38

Change “Telkom” to standard designator

Use ontology-defined location designation

Create fields to represent uncertainty

Step 4: Condition

PWID 2

Ericsson wins deal in Indonesia

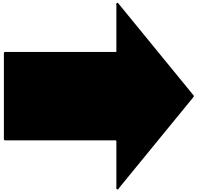
11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers.

The broadband service will include DSL access that will deliver data, voice, and video simultaneously in East Java.

Ericsson's work for this contract is slated to begin in April of 2005.

PT Telkom is Indonesia's largest telecommunications company and currently offers a broadband service, called Speedy, which it launched in July of 2004. State-run Telkom plans to convert its network's infrastructure broadband capacity from fewer than 1,000 lines to 2 million lines by 2008.

Expressions of precision in text



Formal probability distributions

Name	Report 3
OrgAssocType	
Organizations(1) Organization Role:	 Ericsson *Vender*
Organizations (2) Organization role	 PT Telkom *Buyer*
AssocStartDate Lower est. Upper est.	
LocationOfWork: City Providence Country	
Technology	broadband
Contract Amount Estimate Lower bound Upper bound Confidence	 *$7 million* *$5 million* *$9 million* *moderate*
Pedigree Text Processor Source ID	 Ericsson was;;;;; Aerotext, AJJ PWID 2
Region	15, 37

Intermission: Some Manual Editing

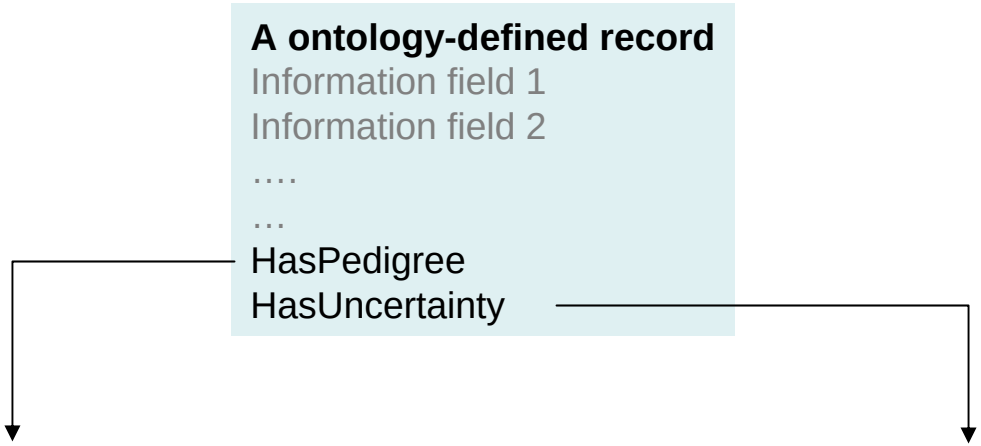
- Review and edit automated system
  - Find sentences text extractor missed
  - Fill ontology-defined fields automated structuring software missed
  - Enter pedigree and uncertainties

PWID 1
Indonesia's Telkom awards Ericsson with broadband contract
Xinhua
11/11/2004 15:01
Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services. Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.
"There is already a very small number of lines—lower than 1, 000—available in Surabaya," PT Ericsson Indonesia President Mitch Lewis was quoted as saying.
He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April.
Telkom launched its broadband service, called Speedy, in July and is aiming to provide 40,000 connections in Jakarta, and 10,000 in Surabaya in the first phase of the project.
It has allocated US$15 million for the service's infrastructure, or between US$290 and US$300 per line.

Manually create a structured record for the missed event

Pedigrees and Uncertainty

Provides qualifying information about data reliability and precision



Pedigree record

- Source
- Date entered
- Method entered
- User-opinion on data credibility
- User initials
- Field identifier

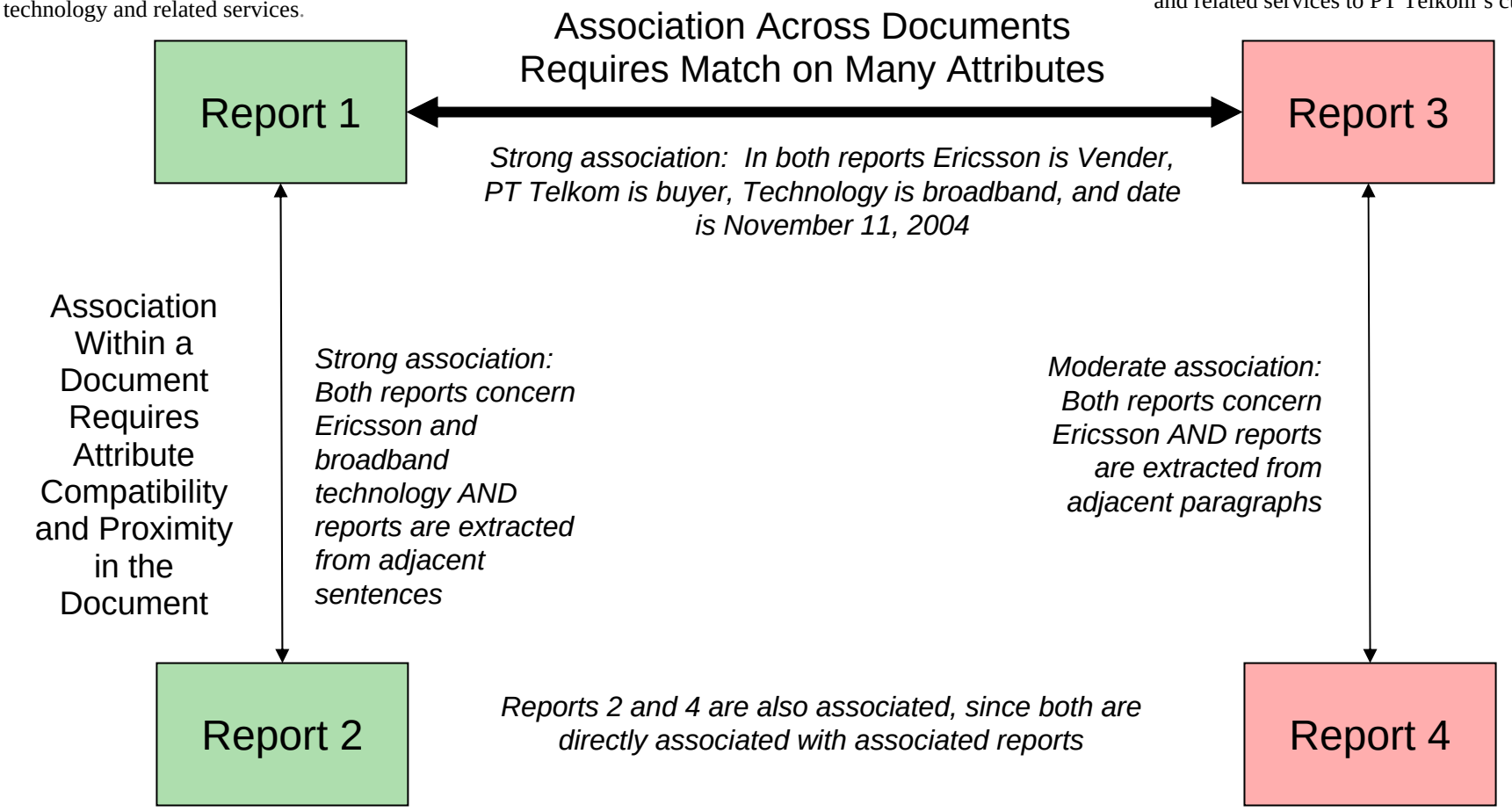
Uncertainty

- Uncertainty representation method
- Uncertainty data (e.g., mean and variance for normal distribution)
- Field identifier

Step 5: Association

Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services.

11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers



Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution

"Missed record" is manually associated

Ericsson's work for this contract is slated to begin in April of 2005.

Step 6: Fuse

Different sentences populate different fields

Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services. (Rpt 1)

Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution (rpt 2)

11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers (rpt 3)
(no info, but provided the link)

Ericsson's work for this contract is slated to begin in April of 2005. (rpt 4)

He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April (missed sentence)

OrgAssocType	Technology Partnership
OrganizationsAndRoles (1)	
Organization	Ericsson
Role:	Vendor
OrganizationsAndRoles (2)	
Organization	PT Telkom
Role:	Buyer
AssocStartDate	
Min estimate	April 1, 2004
Max estimate	April 15, 2004
Confidence	High
LocationOfWork	
City	Surabaya
Region	East Java
Nation	Indonesia
Technology	Broadband
Contract Amount	
Estimate	$7.5 million
Lower bound	$7.5 million
Upper bound	$7.5 million
Confidence	High
Communication Reports	Report 1; Report 2; Report 3, Report 4

Challenges and Research Needs

- Text Extraction
  - Better recall and precision for finding events of interest
  - More complete and accurate represent of text as structured data
- Qualification
  - Formal methods for capturing and representing precision and reliability of text expressions
- Association
  - Metrics for probability of association
  - Techniques (such as multiple hypothesis) for managing uncertain associations

Summary

- Text fusion creates plottable data from unstructured text
- Fusion reduces information overload, integrates Level III products into operational pictures, and supports common situation understanding
- Text fusion processes closely resemble those used in tracking
- Fusion approach
  - Builds on commercial software to find, extract, and structure sentences of interest
  - Augments extractors with fusion-specific software
  - Uses ontology to define data structures
  - Needs some user input